

Modelling Over Dispersed Data due to Excess Zero Using Hidden Markov Model

Ambily Jose^{1*}, Sebastian George²

Department of Statistics, St.Thomas College Pala, Arunapuram, Kerala - 686574, India

ambilystat06@gmail.com, shottom@gmail.com

*Corresponding Author: ambilystat06@gmail.com, Tel.: 9048080975

Abstract— Hidden Markov Models (HMMs) are suitable statistical method for explaining serial dependency in time series data. As per the model, observations are generated from a finite mixture of distributions governed by the principle of Markov chain. Here, we consider Zero Inflated Poisson-Hidden Markov Model (ZIP-HMM) for the series of monthly cholera count data in Kerala during the 2006-2018 period. Cholera is an infectious disease contracted by the consumption of food or water contaminated with Vibrio Cholerae (VC) bacteria. For this analysis, the parameters are estimated by Expectation Maximization(EM) technique. Here, a two-state ZIP-HMM is found to be the better model using Akaike information criterion(AIC) and Bayesian information criterion(BIC). The 'absence' and 'presence' of the VC bacteria are the two hidden states of the fitted model. The best sequence of these hidden states was obtained using Viterbi algorithm.

Keywords: Hidden Markov Model, Overdispersion, Parameter estimation, Vibrio Cholerae, Zero inflated Poisson model.

1. Introduction

The Poisson model is the most commonly used method for modelling time series count data. The equi dispersion feature is the unique principle of Poisson distribution. However in many practical cases, either mean will be greater than variance or vice-versa, resulting in the Poisson assumption becoming wronging real-life count, data are frequently characterised by over dispersion and excess zeros. Zero-Inflated Poisson (ZIP) model provides a way of modelling the excess zeros, besides allowing over dispersion. The ZIP model [1] assumes that the data consist of a mixture of two separate data generation processes, with the first being governed by a binary distribution that generates structural zeros and the second by a Poisson distribution that generates counts, some of which may be zero. The probability mass function (pmf) of ZIP distribution is as given below:

$$\begin{aligned} P(X = x) &= \omega + (1 - \omega)e^{-\lambda}; x = 0 \\ &= (1 - \omega) \frac{e^{-\lambda} \lambda^x}{x!}; x > 0 \end{aligned}$$

where $\omega \in (0, 1)$ and $\lambda > 0$. In the Poisson distribution the mean and the variance are both equal to λ , whereas for $Y \in \text{ZIP}(\omega, \lambda)$:

$$\begin{aligned} E(Y) &= (1 - \omega)\lambda = \mu \\ V(Y) &= \mu + \left(\frac{\omega}{1 - \omega}\right)\mu^2 \end{aligned}$$

A water-born infectious disease that leads to serious watery diarrhea, Cholera results in dehydration and even death, if not treated well. It is caused by the intake of food or water contaminated with a bacteria called Vibrio cholerae (VC). In this paper, we are analyzing the series of monthly cholera

count data in Kerala, India between 2006 and 2018, obtained from the Directorate of Health Services, Government of Kerala. Here, we have two main objectives while applying the model to the data

- Estimating the model parameters from the data by EM technique.
- Decoding the most likely sequence of hidden states by Viterbi algorithm.

The following sections of this paper dealing with, some works related to ZIP-HMM, methodology of HMMs, a real data analysis and describes the results then concludes the research work.

1. Related work

Though several researchers have used ZIP distributions mainly for regression purposes, we can extend their applications to hidden Markov models too. The basic idea of using ZIP in HMM was proposed by Madalina Olteanu and James Ridgway [2], wherein both the Poisson parameter λ and the mixing weight ω depend on the hidden states. Before that DeSantis and Bandyopadhyay [3], Wang [4] also proposed close versions of the same model by introducing a partially observed Markov chain (MC), in which the two components in the ZIP occur according to the states of the partially hidden process. Both articles use the model in a regression context.

2. Methodology

A. Zero Inflated Poisson Hidden Markov model

In an HMM, there is an underlying unobserved state of the system that changes in time according to a Markov process. The distribution of observations at a given time is determined by the system's state at that time [5].

Let $C_t, t \in 1, 2, \dots, T$ be a MC on a finite state space, $S = 1, 2, \dots, m$ with transition probability (TP) matrix $A = (a_{ij})$ where $a_{ij} = P[C_{t+1} = j | C_t = i]$ for any i & $j \in S$ and the initial distribution $\pi = (\pi_1, \pi_2, \dots, \pi_m)$ where $\pi_i = P[C_1 = i]$ for any $i \in S$. The MC C_t , defined on a finite state space is homogenous and irreducible. So the initial distribution π_i is stationary. Hence it satisfies the condition $\pi = \pi A$.

Let $Y_t, t \in N$, a hidden Markov model be a particular type of dependent mixture, with $Y^{(t)}$ and $C^{(t)}$ representing the time from 1 to t , have the following properties.

$$P[C_t | C^{(t-1)}] = P[C_t | C_{t+1}], \quad t = 2, 3, \dots$$

$$P[Y_t | Y^{(t-1)}, C^{(t)}] = P[Y_t | C_t], \quad t \in N$$

Where the first process, $C_t, t = 1, 2, \dots$ satisfies Markov Property; Second one called state-dependent process, $Y_t, t = 1, 2, \dots$ such that the distribution of Y_t depends only on the current state C_t and not on previous states or observations. The conditional probability density function of the observations given the states are given by

$$P[y_1, y_2, \dots, y_T | c_1, c_2, \dots, c_T] = \prod_{t=1}^T p[y_t | c_t]$$

where $p[y_t | c_t]$ is the pmf of the ZIP random variable Y_t when $C_t = i$. Hence for any $t \in N$, consider Z_t an underlying random process such that $Z_t = 1$ leads to a structural zero, with $Z_t | C_t = c_i \sim \text{Ber}(\omega_i)$ [2]. Given the states the conditional pdf of the observations is expressed as under

$$P[Y_t, Z_t | C_t; \theta] = \omega_i^{1_{(1,c_i)}(Z_t, C_t)} (1 - \omega_i)^{1_{(0,c_i)}(Z_t, C_t)} \left(\frac{e^{-\lambda_i} \lambda_i^{Y_t}}{Y_t!} \right)^{1_{(0,c_i)}(Z_t, C_t)}$$

So the model $\{Y_t, C_t\}$, $t = 1, 2, \dots, T$ is said to be ZIP-HMM and is characterised by (i) the stationary initial distribution π , (ii) the TP matrix A and (iii) the state dependent pmfs $P[Y_t | C_t]$. So $\Phi = (\Lambda; A; \Theta)$ denotes the set of parameter space. The parameters to be estimated are the $(m^2 - m)$ transition probabilities (a_{ij}) for any $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, m-1$, the m entries of the vector π , the m parameters of ω_i and λ_i of the m ZIP random variables Y_t . Hence, the vector of unknown parameters is given by

$$\Phi = (a_{11}, \dots, a_{1m-1}, \dots, a_{m1}, \dots, a_{mm-1}, \omega_1, \omega_2, \dots, \omega_m, \lambda_1, \lambda_2, \dots, \lambda_m)$$

which belongs to the parameter space Φ . The initial distribution π will be estimated by the equality $\pi = \pi A$, after the estimation of the matrix A . The estimators of the vector Φ will be obtained by the numerical maximisation of the likelihood by EM algorithm. Then, the likelihood function is given by:

$$L_T = \pi P(y_1) A P(y_2) A P(y_3) \dots A P(y_T) \mathbf{1}'$$

B. Joint Probability Mass Functions of the Model

The HMM $\{Y_t; C_t\}$ with the sequence of observations $y_1, y_2, \dots, y_T$ and the sequence of states of the MC $c_1, c_2, \dots, c_T$, the joint pmf is given by

$$\begin{aligned} P[y_1, y_2, \dots, y_T; c_1, c_2, \dots, c_T] &= \pi_{c_1} a_{c_1 c_2} a_{c_2 c_3} \dots a_{c_{T-1} c_T} p[y_1 | c_1] p[y_2 | c_2] \dots p[y_T | c_T] \\ &= \pi_{c_1} p[y_1 | c_1] \prod_{t=2}^T a_{c_{t-1} c_t} p[y_t | c_t] \end{aligned}$$

Summing over $c_1, c_2, \dots, c_T$ we have the joint pmf

$$P[Y_1, Y_2, \dots, Y_T] = \sum_{c_1} \sum_{c_2} \dots \sum_{c_T} \left\{ \pi_{c_1} p[y_1 | c_1] \prod_{t=2}^T a_{c_{t-1} c_t} p[y_t | c_t] \right\}$$

For an HMM, the real task is to compute the likelihood. Taking into account the large number of sequences in the likelihood, here we are using the Forward-Backward (FB) algorithm instead of the usual computation method. FB is a kind of dynamic programming algorithm that stores intermediate values in a table while building up the probability of the observation sequence. The next task for an HMM is to find the most likely hidden state sequence, which is known as decoding. The most common decoding algorithm for HMM is Viterbi algorithm. The final task for an HMM is to estimate

the parameters of the model, for which EM algorithm [6] is the standard option. EM is an iterative algorithm, which expertly completes the maximisation objective in a two-step procedure.

C. Model Selection

Choosing the right model is an important factor in research. In this study we chose ZIP-HMM for modelling the cholera count. We used AIC and BIC measures for model selection and it is defined as

$$AIC = -2 \log L + 2p.$$

$$BIC = -2 \log L + p \log T.$$

Where $\log L$ is the log-likelihood of the fitted model, p denotes the number of parameters of the model and T is the number of observations. The family which is estimated to be most likely to be true can be identified using the Bayesian approach to model selection and it is known as BIC which differs from AIC in the second term. Compared to AIC, the penalty term of BIC has more weight for $T > e^2$, which holds in most applications. Thus, the BIC often favours models with fewer parameters than does the AIC.

3. Results

We analysed the time series of monthly cholera count data in Kerala during the 2006-2018 period as showed in Figure1. Of the 155 prevalence counts, as many as 106 were noted as zeros, resulting in a scenario where we have excess number of zeros. To be precise, 68% of the total cholera counts in this series is zero. Here, the minimum and maximum values are zero and 14 respectively. While the average value of the series is one, the variance is 4.65. All these point to a clear picture of over dispersion in the cholera count series due to excess zeros.

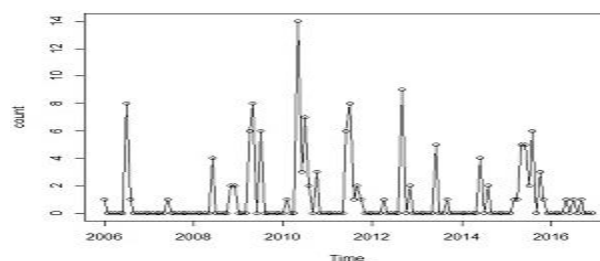


Figure 1. monthly cholera count data in Kerala .

Model parameters are estimated from cholera counts using EM method by R software. Table 1 gives number of parameters m , log-likelihood, AIC and BIC values.

Table 1. ZIP-HMM models fitted to the cholera counts.

M	-Log L	AIC	BIC
1	211.5	428.91	433.09
2	188.79	391.59	412.89
3	181.38	396.77	448.51
4	179.58	421.17	515.52

The estimation of the two-state model was started with (0.5,0.5) as the value of initial distribution π and 0.15 as the off-diagonal values of the transition probability matrix. The plausible starting values of the state-dependent parameters are considered to be $\lambda = (1,2)$ and $\omega = (0.7,0.5)$.

$$\begin{pmatrix} 0.9243 & 0.0767 \\ 0.4952 & 0.5048 \end{pmatrix}$$

After determining the classification of the states of ZIP-HMM, the estimated parameters of two-state model are processed by the Viterbi algorithm [7]. This will bring out the most likely state sequence from which the detected disease counts are generated, as shown in table2. Clearly, this is the hidden sequence of the nature of VC states which resulted in the sudden outbreak of cholera in Kerala.

[illegible]

The value of TP matrix is the probability distribution of transition from one state to another at any time. The TP matrix for all inputs is generally consistent across the study, showing high probability of being in the same state (1 to 1, 2 to 2) and low probability of change from one state to another (1 to 2, 2 to 1). It is clear from TP matrix of two-state ZIP-HMM that 92% of the transition happened between state-one to state-one. This is evident from the state sequence we derived using Viterbi algorithm. These results enable us to conclude that majority of the cholera counts are generated from state 1, wherein the presence of VC bacteria is almost nil.

5. Conclusion

Based on cholera counts in 155 months over a period during the years 2006-2018, we could find that a two-state ZIP-HMM is suitable to describe the stochastic nature of cholera counts per month in Kerala. The states of the MC are absence and presence of VC cholera. We estimated the TP matrix of MC. The two-state ZIP-HMM with the estimated TP matrix can be used to simulate the possible cholera counts per month in coming years. However, one major limitation of these results is that the same pattern of occurrence of cholera may change due to climate change, natural calamities like the recent flood in Kerala. In such situations we have to develop some Markov regression methods for prediction, estimation and simulation in which various covariates should be considered under a Markovian set up. Our further research using Markov regression will be presented latter.

6. References

- [1] Lambert, D. Zero-Inflated Poisson Regression Models with an Application to Defects in Manufacturing Technometrics 1992, Vol.34, pp.1-14.
- [2] Madalina, O.; James, R. Hidden Markov models for time series of counts with excess zeros. European Symposium on Artificial Neural Networks, Belgium, 2012, pp.133-138.
- [3] DeSantis, S.M.; Bandyopadhyay, D. Hidden Markov-models for zero-inflated Poisson counts with an application to substance use. Statistics in Medicine 2011, Vol.30, pp.1678-1694.
- [4] Wang, P. Markov zero-inflated Poisson regression models for a time series of counts with excess zeros. Journal of Applied Statistics 2001, Vol.28, pp.623-632.
- [5] Zucchini, W.; MacDonald, I.L. Hidden Markov Models for Time Series: An Introduction Using R. Chapman and Hall, 2009.
- [6] Baum, L.E.; Petrie, T. Statistical inference on probabilistic functions of finite state Markov chains. Annals of Mathematical Statistics 1966, Vol 37, pp.1554-1563.
- [7] Forney, G.D. The Viterbi algorithm. Proceedings of the IEEE 1973, Vol.61, pp. 268-278.